Nat Lang Semantics (2017) 25:223–261
DOI 10.1007/s11050-017-9135-x

Experiments on the acceptability and possible readings of questions embedded under emotive-factives

Alexandre Cremers[1] · **Emmanuel Chemla**[1]

Published online: 22 March 2017

Abstract Emotive-factive predicates, such as *surprise* or *be happy*, are a source of empirical and theoretical puzzles in the literature on embedded questions. Although they embed *wh*-questions, they seem not to embed *whether*-questions. They have complex interactions with negative polarity items such as *any* or *even*, and they have been argued to preferentially give rise to weakly exhaustive readings with embedded questions (in contrasts with most other verbs, which have been argued to give rise to strongly exhaustive readings). We offer an empirical overview of the situation in three experiments collecting acceptability judgments, monotonicity judgments, and truth-value judgments. The results straightforwardly confirm the special selectional properties of emotive-factive predicates. More interestingly, they reveal the existence of strongly exhaustive readings for *surprise*. The results also suggest that the special properties of emotive-factives cannot be solely explained by their monotonicity profiles, because they were not found to differ from the profiles of other responsive predicates.

Keywords Embedded questions · Emotive-factive predicates · Psycholinguistics

We wish to thank Angelika Kratzer, Andreea Nicolae, Florian Pellet, Yael Sharvit, Benjamin Spector, Kristen Syrett, Lyn Tieu, Wataru Uegaki, anonymous reviewers, and the Attitude Ascriptions & Speech Reports group at SIASSI Berlin. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 313610 and was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

✉ Alexandre Cremers
alexandre.cremers@gmail.com

[1] Ecole Normale Superieure, Paris, France

1 Emotive-factive predicates and questions

Emotive-factive predicates are, as their name indicates, a class of factive predicates of propositional attitudes involving emotions such as surprise, happiness, or regret. In English, although some verbs of this class occur in traditional SVO constructions such as (1a), they also occur in constructions like (1b) and (1c), or adjectival constructions like (1d). Some constructions may also leave the emotion holder implicit, as in (1e). The label 'emotive-factive' thus primarily refers to a semantic property and does not apply to a homogeneous syntactic class of verbs.

(1) a. John regrets breaking the vase.
b. It surprised Peter that Mary came.
c. Peter was surprised that Mary came.
d. Mary is happy that Peter came.
e. It is amazing that John fixed the vase.

Emotive-factive predicates have drawn a lot of attention in the literature on embedded questions, because they are associated with several puzzles that we quickly review below. Given the recent dispute about the possible readings of embedded questions, there now exist some quantitative surveys on this topic, but they only concern predicates which are not emotive-factives (Cremers and Chemla 2016 on *know* and *predict*, Chemla and George 2016 on *agree*, Phillips and George 2016 on *know*, Xiang 2016, Chap. 4 on *know* and *tell*). The goal of this paper is to offer systematic empirical coverage of the behavior of questions embedded under emotive-factive predicates. In this section, we first introduce the phenomena and puzzles of interest. We will observe that judgments and facts are sometimes difficult to assess based on the current claims in the literature.

1.1 Two puzzles regarding questions and emotive-factives

1.1.1 Puzzle 1: Whether-questions

The first puzzle we want to introduce dates back to Karttunen (1977): although emotive-factives generally embed *wh*-questions, they do not embed *whether*-questions, as exemplified by the contrast between (2) and (3a,b) below. Grimshaw (1979) proposed a first analysis of this fact based on the idea that *wh*-clauses embedded under emotive-factives are actually exclamatives, but recent theories tend to treat them as genuine questions and to put forward various other factors to explain the ungrammaticality of (3a,b). Some attribute it to the competition between the embedded questions in (3a,b) and the embedded *that*-clauses in (4a,b), which are made equivalent by a presupposition stronger than factivity ('speaker-factivity' or 'super-factivity': see Guerzoni and Sharvit 2007; Guerzoni 2007; Sæbø 2007) or by anaphoric properties of emotive-factives (Herbstritt 2014; Roelofsen et al. to appear). Other analyses propose that questions come with a determinate strength of exhaustivity and that *whether*-questions force a strongly exhaustive reading which is incompatible with

emotive-factives (see Nicolae 2015; Guerzoni and Sharvit 2014). Abels (2007) proposes that the polar question in (3a) systematically fails to satisfy the presupposition of *surprise* (but does not account for (3b)). Finally, Romero (in press) proposes that the focus sensitivity properties of emotive-factives are at the source of the deviance of (3a,b).

(2) It is amazing what they serve for breakfast.

(3) a. * It is amazing whether they serve breakfast.
b. * It is amazing whether they serve TEA or COFFEE for breakfast.

(4) a. It is amazing that they serve breakfast / that they do not serve breakfast.
b. It is amazing that they serve tea / that they serve coffee for breakfast.

The different proposals also differ on the characterization of the puzzle itself. While most authors take the unacceptability of (3a,b) as a case of plain ungrammaticality, Sæbø (2007) argues, based on examples like (5), that under some circumstances *whether*-questions are in fact acceptable with emotive-factives. Meanwhile, Herbstritt (2014) and Roelofsen et al. (to appear), unlike others, attribute the ungrammaticality of the polar question (3a) and the alternative question (3b) to different mechanisms, which might lead us to expect differences in acceptability judgments for the corresponding sentences.

(5) Don't read this installment before seeing the episode if you want to be surprised at whether or not Hercules makes it.

1.1.2 Puzzle 2: Exhaustive readings

Focusing on well-formed constructions, we note that another dispute concerns the possible meanings of constructions involving emotive-factives. Several readings have been proposed for sentences with questions embedded under verbs like *know*, as in (6). Karttunen (1977) first proposed the reading in (6a), which was later named "weakly exhaustive" (WE), while Groenendijk and Stokhof (1982, 1984) argued for a stronger reading, namely the "strongly exhaustive" (SE) reading in (6b).

(6) Mary knows who was at the party.
a. For each person who was at the party, Mary knows that he/she was.
b. For each person who was at the party, Mary knows that he/she was and she knows that no one else was.

While *know* was traditionally considered to only (or predominantly) receive SE readings (Groenendijk and Stokhof 1982, 1984), Berman (1991) argued that *surprise* must convey a WE reading. This introspective judgment has been endorsed by most later authors—notably Heim (1994), who took it as the main argument for a theory in which embedded questions are ambiguous between an SE and a WE reading.[1] A

[1] In Heim's (1994) theory, full sentences with embedded questions are *not* ambiguous, because the predicate selects the appropriate reading of the question it embeds.

crucial empirical difference between *know* and *surprise* would be that, in a situation where Mary knows who the students are, the inference in (7) seems valid while the inference in (8) does not. Since the SE but not the WE reading makes the two questions "which students called" and "which students didn't call" equivalent, we may conclude that *know* but not *surprise* receives an SE reading.

(7) Mary knows which students called.
⇒ Mary knows which students didn't call.

(8) It surprised Mary which students called.
⇏ It surprised Mary which students didn't call.

1.2 Monotonicity as a key to puzzles 1 and 2?

It has been suggested that some of the puzzling facts related to questions under emotive-factive predicates may be explained by their specific entailment patterns (Lahiri 2002, p37; Guerzoni 2003, 2007; Guerzoni and Sharvit 2007; Uegaki 2015), which were first studied by Wilkinson (1996) to explain the interaction between emotive-factives and negative polarity items (NPIs). As an example, some theories derive SE readings of embedded questions as implicatures (following Klinedinst and Rothschild 2011), and therefore predict that they should not arise in the scope of certain environments where scalar implicatures do not arise, such as negation and some other downward-entailing operators.[2] *Whether*-questions also have specific interactions with NPIs, and some accounts relate the distribution of NPIs and *whether*-questions under responsive predicates (Guerzoni 2003; Guerzoni and Sharvit 2007, 2014).

However, the different studies diverge regarding the monotonicity they attribute to *surprise* (downward-entailing or non-monotonic), and little has been said about other emotive-factive predicates (although see Uegaki 2015 on *be happy*). We therefore tested the Strawson-entailment patterns associated with some of these predicates in the same way that we collected data for their selectional properties and possible readings; they are interesting because (a) monotonicity plays a role in some popular accounts, (b) experimental data would help arbitrate between conflicting introspective judgments,[3] and (c) the experimental literature suggests that 'perceived monotonicity' may be the relevant factor (Chemla et al. 2011) and that it differs from actual monotonicity (e.g., because downward monotonicity is poorly evaluated in experimental tasks; Geurts and van der Slik 2005).

[2] In fact, Klinedinst and Rothschild (2011) argue that *surprise* does have an SE reading, the derivation of which would require a local application of the exhaustivity operator in a downward-entailing environment. Although very close in spirit, Uegaki (2015) departs from this view in assuming that this reading is not available and by deriving its unavailability from the non-monotonicity of *surprise*.

[3] Note that the theoretical characterization of the monotonicity of a predicate depends crucially on its denotation. Uncertainty about the monotonicity of emotive-factive predicates amounts essentially to uncertainty about their denotations. Experimental explorations of the semantic properties of a predicate may thus inform theoretical semantics by constraining conceivable denotations.

1.3 Summary

Given the interest in emotive-factive predicates and the large variety of theories proposed to account for their properties, the main goal of this project was to gather quantitative data in a theory-neutral perspective (as much as possible). We tested which types of questions can be embedded under emotive-factives, what (exhaustive) readings these embedded questions would carry, and what monotonicity properties are associated with emotive-factives. We focused on testing various widespread claims from the literature rather than predictions of specific theories, but this will not prevent us from pointing out certain precise theoretical consequences of our data along the way.

2 Experiment 1: Selectional properties of attitude predicates

2.1 Goals

The primary goal of this experiment was to test the selectional properties of emotive-factives, but, as we explain below, this experiment was also an opportunity to address a few other empirical questions regarding the selectional properties of various other attitude predicates.

2.1.1 Selectional properties

The main goal of this experiment was to test the unacceptability of *whether*-questions under emotive-factive verbs (Karttunen 1977; Grimshaw 1979). We also tested the selectional properties of other attitude predicates and compared different complements. In Lahiri's (2002) typology, emotive-factive predicates fall into the category of *responsive* predicates, which take both declarative and interrogative complements (e.g., *know*), in contrast with *rogative* verbs, which only embed interrogatives (e.g., *wonder*), and *anti-rogative* verbs, which only embed declaratives (e.g., *believe*).[4] We tested the (un)acceptability of *whether*-questions under emotive-factives against baselines of perfectly acceptable constructions and of constructions where the selectional properties of the embedding predicate are violated. We also included the verb *regret*, which has been claimed to be an exception among emotive-factives for not embedding questions at all.

2.1.2 Degrees of unacceptability and the nature of the constraints

We will present the results of a task in which participants are asked to report their acceptability judgments on a continuous scale (more on this below). We will not make claims about how participants may have understood the task and we will not draw conclusions from absolute judgments that people report on this scale. But the scale will

[4] We use the terminology from Theiler et al. (2016), since anti-rogative verbs are not discussed in Lahiri (2002).

be calibrated by collecting baselines with allegedly acceptable sentences (questions embedded under rogative predicates and declarative complements embedded under anti-rogative predicates) and unacceptable sentences (the opposite: questions embedded under anti-rogative predicates and declarative complements embedded under rogative predicates). If these judgments come out right, we will be content to suppose that the type of judgments we collect are of relevance to our linguistic inquiry and will be in a position to interpret responses to more controversial cases relative to these established baselines.

To take a different perspective on this practical issue: gathering quantitative acceptability measures of this sort may also allow us to evaluate finer-grained differences in the degree of (un)grammaticality of different constructions; some violations may have a stronger impact than others. Even without assuming differences in the degrees of ungrammaticality, given the specifics of our experimental design we may be able to detect some types of violations but not others. Let us be more concrete and see how this could have theoretical relevance. Guerzoni (2007) and Sæbø (2007) propose that *whether*-questions are ruled out under emotive-factives because they compete with declaratives at the pragmatic level. In contrast, Nicolae (2013, 2015) and Guerzoni and Sharvit (2014) propose that *whether*-questions are encoded as strongly exhaustive and that emotive-factives select weakly exhaustive questions only. The former view predicts mere pragmatic infelicity for *whether*-questions under *surprise*, while the latter predicts a stronger grammatical incompatibility. In fact, the approach proposed in Herbstritt (2014) and Roelofsen et al. (to appear) even predicts a difference between polar questions and alternative questions; the infelicity of alternative questions under emotive-factives is derived from a contradictory presupposition, whereas polar questions are eliminated through competition with a declarative complement (using the maxim of manner; Grice 1975).[5] We may expect all these differences to translate into different degrees of unacceptability, even though the link between the strength of incompatibilities and their nature (syntactic, semantic, or pragmatic) is not entirely clear.

2.1.3 Quantificational variability

Quantificational variability effects (QVE; Berman 1991) constitute another phenomenon related to embedded questions which has given rise to conflicting judgments in the literature. The effect is visible in (9), which has a reading that is true if and only if Mary knows of most students who called that they called.

(9) For the most part, Mary knows which students called.

The availability of QVE has been debated for rogative verbs. On the one hand, Lahiri (2002) argues that rogatives can never receive QVE readings; his theory predicts a semantic type mismatch in any possible structure. Beck and Sharvit (2002),

[5] However, Roelofsen et al. (to appear, fn. 13) discuss the possibility that this infelicity has been grammaticalized over time. One difference between their account and Guerzoni (2007) is that under emotive-factives they predict perfect synonymy between the polar question *whether-p* and the *positive* declarative *that-p*, whereas the competitor of the question in Guerzoni (2007) depends on which of p or $\neg p$ is true in the world of evaluation.

on the other hand, argue that in some cases QVE readings may be available. They propose a semantics which can derive QVE for any rogative verb and suggest that the unavailability of this reading in Lahiri's (2002) examples is due to an independent, softer constraint (subject to contextual variation). We took our experiment as an opportunity to gather data on this issue. The idea was to use the acceptability of sentences where a predicate modified by a quantity adverb embeds a *wh*-question as a proxy for the availability of QVE. We did not test which reading participants had for the target sentences, but our intuition was that if QVE readings are available, these sentences should be more natural and receive a higher rating. Of course, many other factors may affect the acceptability of these adverbs and one should keep in mind that the measure of their acceptability is only a proxy for the availability of QVE. Nevertheless, foreshadowing the results of our experiment, the fact that the distribution of acceptability ratings perfectly matched the judgments reported in Beck and Sharvit (2002) regarding the availability of QVE makes this factor a good candidate to explain the differences we observed.[6]

2.2 Methods

2.2.1 Task and instructions

Participants were asked to provide acceptability judgments for different sentences. They provided a continuous response with a horizontal slider the ends of which were labelled 'weird' and 'natural', as illustrated in Fig. 1.

The instructions introduced a common context for the sentences. All sentences were about aliens visiting the Earth, so that plausibility or conflict with world knowledge would not interfere with participants' judgments. To encourage participants to judge the grammaticality of the sentences, and in particular to focus on selectional properties of the verbs, the instructions provided the two examples in (10), which were described as natural and odd respectively.

(10) a. Peter saw a fluffy purple alien playing the piano.
b. Peter went a fluffy purple alien playing the piano.

Four items were presented as warm-up to the participants, before the experimental phase. These sentences were comparable to the sentences presented in the instructions and of no interest to us. Nothing distinguished the warm-up phase and the experimental phase.

2.2.2 Design and stimuli

Two factors were crossed: Embedding Predicate (17 predicates from four broad categories, see Table 1) and Complement Type (5 levels, see Table 2). Three instances of each combination were generated, for a total of 255 experimental items.

[6] Lahiri (2002) shows that another reading may be available for rogative verbs with quantity adverbs, the 'focus-affected' reading, but this reading is intuitively harder to get and may require a specific prosody. In any case, the focus-affected reading should be equally available for all rogative and responsive predicates, so any difference we would find between the two types of predicates cannot be attributed to this reading.

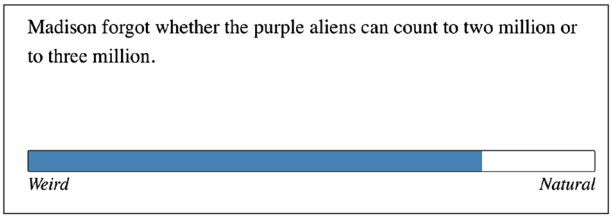

Fig. 1 Example of an item from Experiment 1, with an alternative question under *forget*

Each item was generated by drawing randomly from lists of *proper names* for the agents, *adjectives* characterizing the aliens, and *predicates* describing what the aliens did (complete lists are presented in Appendix 1). For each complement type, the embedded clause (interrogative or declarative) was inserted in the COMP argument position of the schematic structures in Table 1. In the case of Adv+*Which*-Question, the *which*-question was in the COMP position but the adverbial phrase was appended in sentence-initial position. Polar Question and Alternative Question only differed

Table 1 List of embedding predicates tested in Experiment 1 and the associated sentence structures

Predicate type	Predicate	Schematic structure
Emotive-factive	*Surprise*$_1$ (double object)	It surprised X COMP.
	Surprise$_2$ (passive)	X is surprised (by) COMP.
	Regret	X regrets COMP.
	Be happy	X is happy (about) COMP.
Responsive	*Know*	X knows COMP.
	Remember	X remembers COMP.
	Forget	X forgot COMP.
	Misunderstand	X misunderstands COMP.
	Disregard	X disregarded COMP.
	Agree	X_1 and X_2 agree (on) COMP.
	Guess	X guessed COMP.
Anti-rogative	*Believe*	X believes COMP.
	Think	X thinks COMP.
	Assume	X assumes COMP.
Rogative	*Wonder*	X wonders COMP.
	Depend	COMP depends on genetics.
	Ask	X asked COMP.

Prepositions in parentheses were present with interrogative complements but absent with declarative complements. *Depend* did not require any proper name; *agree* required two names, and we ensured that these two names would always be different.

Table 2 List of complement types tested in Experiment 1

See text for details.

Type	Structure
Declarative	…that the ADJ aliens PRED.
Polar question	…whether the ADJ aliens PRED.
Alternative question	…whether the ADJ aliens PRED or PRED′.
Which-question	…which aliens PRED.
Adv+*Which*-question	For the most part, …which aliens PRED.

in their PRED complement: a specific version of each PRED with a disjunction was used to generate alternative questions. Note that questions of the form 'whether A or B' are inherently ambiguous between an alternative and a polar reading (and this is particularly true when they are presented as written stimuli rather than audio). To avoid such confusion, we worked with native speakers to obtain questions which would be as biased as possible towards the alternative reading, but we cannot guarantee that these test items were never interpreted as polar questions (we will come back to this in the presentation of the results). Examples for each complement type embedded under various predicates are given in (11)–(15).

(11) *surprise*$_1$ + Declarative:
"It surprised Mary that the fluffy aliens play the piano with their wings."

(12) *wonder* + Polar Question:
"Peter wonders whether the hollow aliens can eat 5 pounds of licorice."

(13) *know* + Alternative Question:
"Grace wonders whether the red aliens drink soda with a straw or with a spoon."

(14) *agree* + *Which*-Question:
"Alex and Madison agree on which aliens write poems about the moon."

(15) *believe* + Adv + *Which*-Question:
"For the most part, Jacob believes which aliens ride tall purple horses."

2.2.3 Participants

In all, 50 participants were recruited on Amazon's Mechanical Turk and were paid $2 for their participation (age range: 21–64, 29 males).

2.2.4 Statistical methods

The data and analysis script for each experiment are available online at http://semanticsarchive.net/Archive/GRhZmM4N/Cremers-Chemla-ExpEmotiveFactives.html. Results were analyzed with `R` and the `lme4` package (R Core Team 2014; Bates et al. 2015b) and plotted with `ggplot2` (Wickham 2009).

To compensate for differences in how participants used the continuous scale, responses were centered and normalized by participant before any analysis. We followed the procedure proposed by Bates et al. (2015a), who suggest to first try to fit maximum random effects structures by participants and by item (as suggested in Barr et al. 2013) and then eliminate useless components in the random effects structure to avoid over-parametrization.[7]

[7] More precisely, we always fitted models with maximal by-subject and by-item random structures, although they often did not converge. Then we used the function `rePCA` provided in the package `RePsychLing` to remove all components which explained less than 0.1% of the variance explained by the main component. This was done independently for the subject and item random effects structures, because random effects associated with items tend to be smaller. We also removed correlations between random effects in the updated "parsimonious" model. A second pass ensured that no component under the threshold remained in the new model (this could happen when the PCA had been done on a maximal model which had not

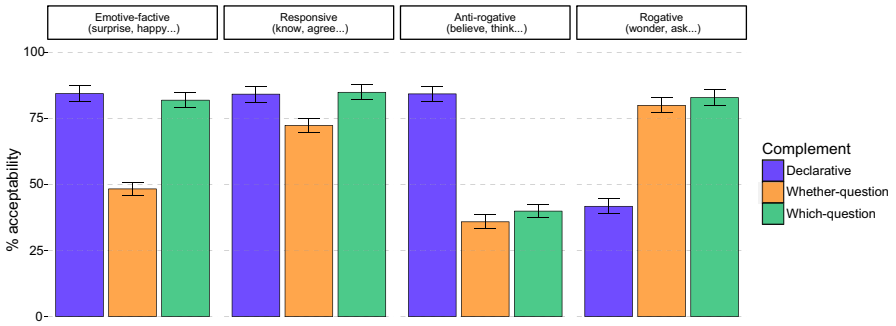

Fig. 2 Aggregated acceptability by predicate category and complement category in Experiment 1. In this graph and all others, error bars correspond to standard error of the mean. The verbs *regret*, *misunderstand*, and *disregard* were excluded from the analyses (see Fig. 3 for details). *Whether*-questions include both polar and alternative questions. Sentences with an adverb of quantity are not shown in this figure.

For each relevant parameter, we report the 95% confidence interval (CI) rescaled to the original 0–100 scale used in the graphs. The theoretically relevant comparisons all yield very clear results, as the graphs will show.

2.3 Results

Figure 2 presents a summary of the results by predicate type (see Table 1). For a break-down by predicate and results from the sentences with an adverb of quantity, see Fig. 3. In Sect. 2.3.1, we first discuss the different categories of predicates one by one, ignoring sentences with adverbs of quantity and differences between the two types of *whether*-questions. We turn to the more specific issues addressed by these conditions in a second stage of analysis in Sect. 2.3.2.

Misunderstand, *disregard*, and *regret* were removed from the analyses, since these predicates did not reach 75% acceptability in any condition.

2.3.1 Predicate categories

Under anti-rogative predicates, declarative complements were well-accepted (CI: [79, 86]), whereas interrogative complements were degraded (CI: [34, 43] for *whether*-questions and [38, 48] for *which*-questions).[8]

Under rogative predicates, we observed the opposite pattern. Declarative complements were clearly degraded compared to all interrogative complements (CIs: [39, 49]

Footnote 7 continued
converged). All parsimonious models converged without requiring any further simplification, and whenever the maximal models had converged, we found that they did not explain more variance than the parsimonious models.

[8] On declarative sentences, there was no difference between *believe* ([82, 88]), *think* ([79, 88]), and *assume* ([75, 85]). With *wh*-questions, *assume* was rated higher ([55, 69]). A quick search on the internet indeed revealed some examples of questions under *assume*, such as "People like to assume which of my parents is black or which of my parents is white."

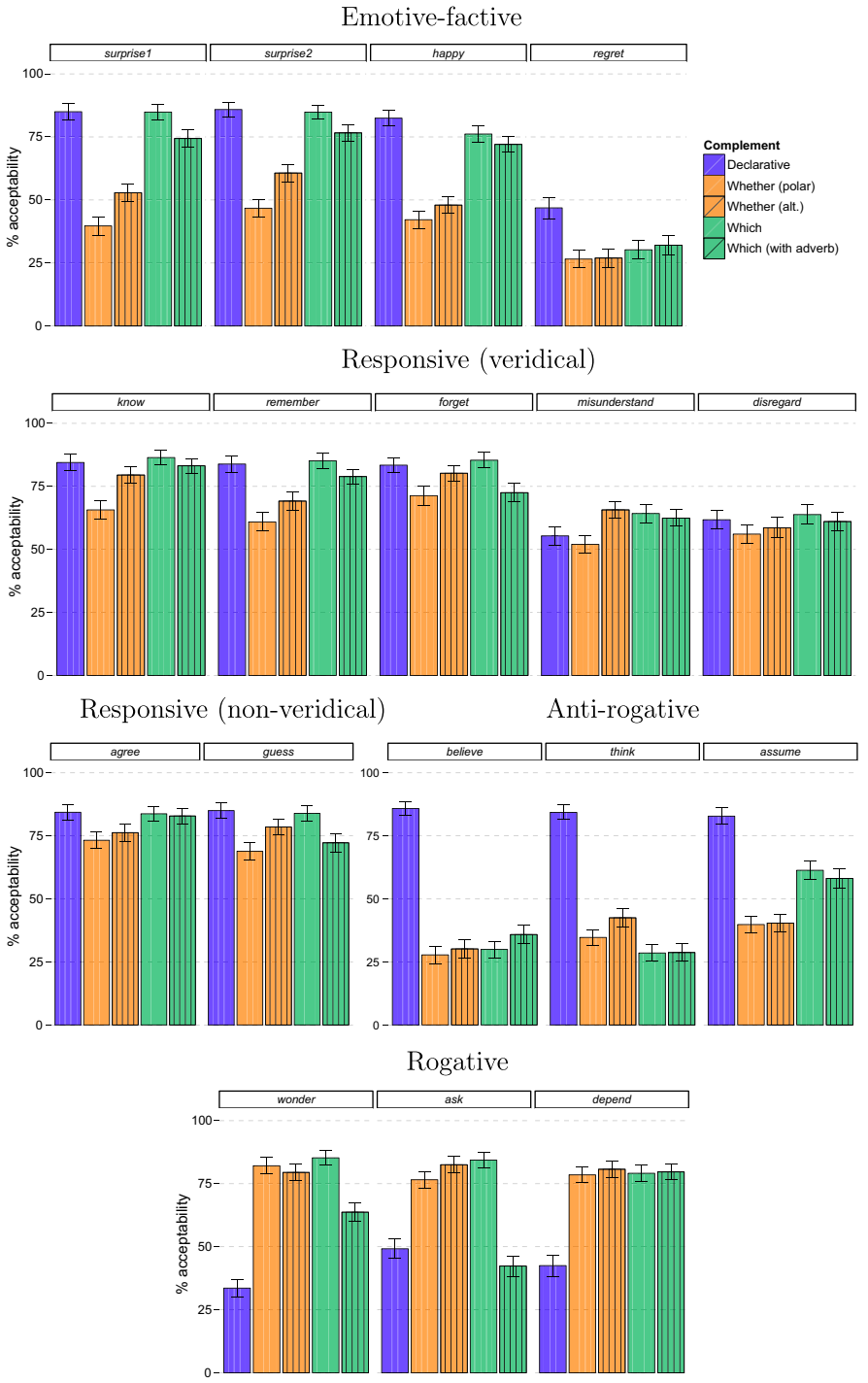

Fig. 3 Acceptability for each predicate and complement in Experiment 1, by predicate category

vs. [75, 81]). This difference showed up for all three verbs we tested, although the contrast was slightly reduced for *ask*.

With responsive predicates, all complements were clearly above 50%, although *whether*-questions were slightly degraded. Leaving aside *misunderstand* and *disregard*, all these verbs were qualitatively similar: equally compatible with declarative complements ([79, 85]) and *wh*-questions ([80, 86]), and almost as acceptable with *whether*-questions ([69, 74]). Veridical (*know, forget, remember*) and non-veridical (*agree, guess*) responsive predicates did not differ significantly: a model with veridicality, complement type, and their interactions as fixed effects showed no effect of veridicality and no interaction (all $t < 1.5$, estimated effects below 1% acceptability).

Under emotive-factive predicates (excluding *regret*), declarative and *which*-questions were as acceptable as under responsive verbs ([79, 86] and [78, 83] respectively).[9] This is in contrast with *whether*-questions, which were significantly degraded compared to their acceptability under other responsive predicates ([46, 54] vs. [69, 74]) but still more acceptable than under anti-rogative predicates ([34, 43]). There was no significant difference between the overall acceptability of $surprise_1$, $surprise_2$, and *be happy*. *Regret* was very degraded, even with declarative complements.

2.3.2 Specific questions

Rogative predicates and QVE: Under responsive, anti-rogative, and emotive-factive predicates, we found that the presence of an adverb slightly reduced the acceptability of a sentence with an embedded *wh*-question ([64, 67] vs. [68, 71]). The effect of the adverb did not interact with the verb category ($\chi^2(4) = 5.0$, $p = .29$) but it had an important effect within the class of rogative predicates: sentences with adverbs of quantity were judged as unacceptable as sentences with declaratives for *ask* ([36, 52]), as acceptable as sentences with interrogatives for *depend* ([75, 82]), and somewhere in between for *wonder* ([58, 69]).

Polar and alternative questions: As is visible in Fig. 3, polar questions were less acceptable than alternative questions across all predicates ([62, 67] vs. [69, 73]). First, as mentioned earlier, alternative questions could in principle be interpreted as polar questions. The fact that we observed a difference suggests that this was not always the case, but one should keep in mind that actual contrasts between polar and alternative questions may be stronger than the effect we observed. This difference may seem surprising, because alternative questions were strictly more complex than polar questions (they only differed in that they had an extra disjunct appended). From discussion with a dozen informants, it turned out that some speakers, mostly from western regions of the USA, dislike embedded *whether*-questions without an overt disjunction (they would need to add "or not" in polar questions). With 50 participants, from only 26 states in the USA, and no precise information regarding their linguistic background, we did not have the adequate resources for a proper study of American dialects. Nevertheless, longitude was a significant predictor of individual differences in the acceptability of polar and alternative questions ($t = -2.5$, $p = .02$). Latitude did not have a signifi-

[9] There is actually a statistically significant difference, but it is qualitatively very small: [0.8, 3].

cant effect, but there was a trend for an interaction ($t = 1.7$, $p = .09$).[10] We would like to point out that Guerzoni and Sharvit (2014) assume that polar questions are underlyingly alternative questions with a silent "or not". The fact that some dialects of American English require an overt "or not" suggests that such an analysis of polar questions is on the right track.

2.4 Discussion

Our results confirm most judgments reported in the literature (see also Sprouse and Almeida 2013; Sprouse et al. 2013 on the reliability of introspective judgments): the four classes of predicates we discussed correspond to a natural typology based on acceptability of declarative and interrogative complements. As expected, the anti-rogative predicates were bad with any interrogative complement, the rogative predicates displayed the opposite pattern (good with interrogatives, bad with declaratives), and the responsive predicates were good with any complement. Emotive-factive predicates were equally good with declarative and *wh*-questions, but clearly degraded with *whether*-questions.

Turning to new conclusions we can draw from this experiment, we observe that *whether*-questions under emotive-factives, although degraded, are not as unacceptable as under anti-rogative predicates. This suggests that the constraint ruling out *whether*-questions under *surprise* must be different from the one ruling out questions under *believe* (which itself is not well explained). Under the simplistic assumption that pragmatic constraints are 'softer', or more concretely that in our task participants do not judge pragmatic violations too low, this result could support pragmatic accounts such as Guerzoni (2007) and Sæbø (2007) for the unacceptability of *whether*-questions under emotive-factives. These authors argue that the unacceptability is due to a contextual competition with *that*-clauses (note however that Sæbø proposes a similar argument for questions under *believe*, which is not supported by our results). Nevertheless, one should be very careful in interpreting the strength of a constraint as an indication of its semantic or pragmatic nature.

We did observe a contrast between polar and alternative questions, but it was not specific to emotive-factive predicates and, if anything, it went in the opposite direction from what Herbstritt (2014) and Roelofsen et al. (to appear) might predict. Most likely, this contrast can be explained as a syntactic constraint on the construction of polar questions in some western dialects of American English. By contrast, we did not observe any difference between the two constructions for *surprise* (the impersonal construction and the passive-like construction with the presupposition *by*).

Regret was degraded with interrogatives, but also with declarative complements. The first fact is a known puzzle in the literature (Lahiri 2002; Egré 2008); the second

[10] We would like to thank Aparicio Kozuch, who insisted on adding "or not" to polar questions when we were eliciting judgments for another project, thus suggesting this hypothesis. We would also like to thank the many American linguists we met during the summer 2015 who kindly provided judgments, and Florian Pellet, who included the question about state when programming the experiment and thus allowed us to test the hypothesis. One outlier from Indiana who systematically rejected polar questions had to be removed from the analyses.

one is more surprising. It may be that *regret* requires some relation between its agent and the proposition denoted by its complement. In our stimuli, the propositions denoted by the declarative complements were facts about aliens, completely independent from the agents. Alternatively, the class of emotive-factives may not be as homogeneous as previously assumed and perhaps should be split into subclasses, with predicates like *surprise* and *be happy* distinguished from verbs like *regret* and *resent*, which do not embed questions.

Finally, the results for sentences headed by *for the most part* suggest variation in the availability of QVE for rogative verbs. Of course, since we did not probe what readings participants accessed but only the acceptability of the test sentences, we cannot be sure that the availability of QVE is what drives the differences. However, we would like to point out that (a) QVE seems to be the most salient reading for these sentences (the alternative would be what Lahiri (2002) calls the focus-affected reading, which would not explain the differences between predicates), and (b) the differences we observed perfectly match Beck and Sharvit's (2002) judgments on the availability of QVE (available with *depend on*, unavailable with *ask*, unclear and possibly context-dependent with *wonder*).

Before addressing the question of strong exhaustivity, we will first investigate the monotonicity of emotive-factive predicates. This will inform us about the semantics of the predicates themselves, without which we cannot even be sure of what their exhaustive readings are.

2.5 Conclusions for Experiment 1

We collected naive speakers judgments pertaining to the selectional properties of embedding predicates. The results confirmed most judgments from the theoretical literature. New observations include the contrast between the acceptability of *whether*-questions under emotive-factive and anti-rogative predicates, a small difference between polar and alternative questions (which may be explained by dialectal differences), and data on adverbs of quantity which may inform the debate on QVE with rogative predicates. In the remainder of this paper, we move on to more semantically oriented investigations, focusing on constructions that were judged grammatical in this experiment. We will first investigate the monotonicity of emotive-factive predicates when they embed straight declarative complements. This will inform us about the semantics of the predicates themselves and is in fact a prerequisite to studying exhaustive readings they could give rise to when they embed questions, an issue we turn to in the final experiment.

3 Experiment 2: On the monotonicity of responsive predicates

3.1 Goal

The goal of this experiment was to determine the monotonicity of some emotive-factive predicates. This is important for at least two reasons. First, various special properties of these verbs (e.g., NPI licensing abilities, selectional properties, avail-

able readings) have been linked to their monotonicity profiles. Second, judgments in the theoretical literature diverge. Guerzoni and Sharvit (2007) consider *surprise* to be Strawson-downward entailing, whereas Uegaki (2015) takes *surprise* to be Strawson-non-monotonic. Furthermore, most of the literature focuses on *surprise*, and we thought it would be interesting to test *be happy* which, we hypothesized, might be judged 'less' downward entailing.

It is indeed worth noting that the relevant theoretical concept is not classical monotonicity but Strawson-monotonicity, which describes the entailment patterns between complements of a predicate, provided all presuppositions are satisfied. In order to assess Strawson-monotonicity, we thus tested monotonicity-like inference patterns, augmented with premises which would guarantee that the presuppositions of all relevant propositions would be true.

3.2 Methods

3.2.1 Task and instructions

This experiment was an inferential task with continuous responses, in which we tested the inference patterns of various environments that embed propositions. While graded responses are common in experiments collecting grammaticality judgments, their use in an inferential task may be a non-obvious decision. We submit that similar arguments can hold for their use in both cases, however: they may help detect otherwise hidden effects, and differences they reveal call for an explanation. Furthermore, we hope that testing systematically both upward and downward inferences for a single environment can help clean up some of the artificial effects which may have been created by the response scale. Other experiments using such a response scale have convinced us that they could yield fruitful and relevant results in line with the kind of judgments linguists typically work with (a similar scale was applied to collect judgments about presuppositions in Chemla 2009, scalar implicatures in Chemla and Spector 2011, as well as monotonicity judgments in Chemla et al. 2011). As illustrated in Fig. 4, experimental items consisted of two premises and one conclusion.

The instructions introduced a minimal common context for the whole experiment, as in Experiment 1 (all sentences were about aliens who have just spent a week on

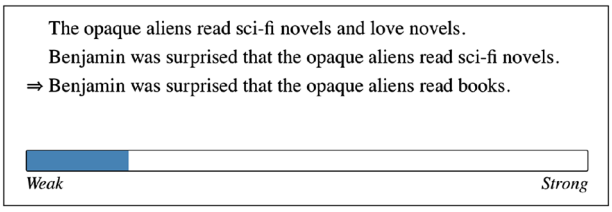

Fig. 4 Example of a direct item in Experiment 2. One can see that the predicate in the second premise (*read sci-fi novels*) is stronger than the predicate in the conclusion (*read books*). The "fact" (first premise) ensures that the opaque aliens read sci-fi novels, thus validating the factive presuppositions of both the second premise and the conclusion.

Earth). The instructions provided the two examples in (16a)/(17a) followed by the explanations in (16b)/17b).

(16) a. There are pink aliens and blue aliens.
Less than 50 pink aliens play the piano.
⇒ Less than 50 pink aliens play the piano with their left hand.

b. "In this case, you should put the bar close to the right *[strong]*. Indeed, if in total less than 50 pink aliens play the piano, there cannot be more than 50 pink aliens who play the piano with their left hand."

(17) a. The blue aliens live in tree houses 10 feet above the ground with rope ladders.
Peter knows that the blue aliens live in tree houses.
⇒ Peter knows that the aliens live in tree houses 10 feet above the ground.

b. "Here you should put the bar close to the left *[weak]*, because even if Peter knows that the blue aliens live in tree houses, nothing indicates that he knows that their tree houses are 10 feet above the ground."

The two examples in the instructions were given as the first actual items, so that participants would be familiarized with the presentation and task before tackling actual items of interest.

3.2.2 Design and stimuli

We tested direct inferences, where the proposition embedded in the premise entailed the proposition embedded in the conclusion, and indirect inferences, where the proposition embedded in the second premise was asymmetrically entailed by the proposition embedded in the conclusion. We crossed two factors: Environment (8 levels) and Direction of the Inference (2 levels). Participants saw 8 repetitions of each combination plus 16 filler items, for a total of 144 items.

Most sentences tested in this experiment were minor modifications of sentences which received very high ratings in Experiment 1.

Participants were asked to indicate how strongly the conclusion followed from the premises using a horizontal slider whose ends were labelled 'weak' and 'strong'. The first premise was always a "fact" which ensured that the presuppositions of the second premise and the conclusion were satisfied, thus allowing us to test Strawson-entailment rather than classical entailment. To make the sequence less repetitive and more natural, the fact was always a conjunctive statement, with one conjunct being the presupposition (of the second premise or of the conclusion, depending on which one was the strongest). The second premise and the conclusion only differed in the predicate of their embedded clause. These predicates could be "strong" (PRED_+) or "weak" (PRED_-), and this allowed us to test both directions of inference. Direct items tested the inference from the strong to the weak, which is only valid in Strawson-upward-entailing (SUE) contexts. Indirect items tested the inference from the weak to the strong, which is only valid in Strawson-downward-entailing (SDE) contexts. They were obtained by switching the conclusion with the second premise in a Direct

Table 3 List of attitudes tested in Experiment 2, and the associated sentence structures

Predicate type	Predicate	Schematic structure
Emotive-factive	*Surprise*	X was surprised that…
	Be happy	X was happy that…
Other responsive	*Know*	X knew that…
	Forget	X forgot that…
	Agree	X_1 agrees with X_2 that…

Agree required two names, which we ensured were always different from each other. We used the *agree with* form this time because it introduces a presupposition without being a factive verb.

item. Strawson-non-monotonic (SNM) contexts make both the direct and the indirect inferences invalid.

Table 3 presents the attitude predicates tested in Experiment 2, and Table 4 presents the structure for each item type. We tested the emotive-factives *surprise* and *be happy*. To keep the sentences with emotive-factives superficially similar to those using other attitude predicates, we used the SVO-like construction: 'Mary is surprised/happy that…'. We compared *surprise* and *be happy* to three other responsive verbs: *know* and *agree*, which we expected to be SUE, and *forget*, which we expected to be SDE. For *agree*, we used the structure 'Peter agrees with Mary that…'.[11] We used three control constructions without attitude verbs to serve as a baseline for UE, DE, and NM environments.[12] We created valid fillers to counterbalance the fact that NM controls made all inferences invalid. All items were in the past tense, except for *agree*, where the main verb was in present tense.

Each participant saw 8 repetitions of each condition, obtained by picking a triplet of instantiations for PRED, ADJ, and proper names, with the constraint that Direct items and corresponding Indirect items were matched for each participant (i.e., constructed on the same 8 triplets). Each participant thus saw a total of 144 items: 8 (environment) × 2 (direct vs. indirect) × 8 (repetitions) test items, and 16 direct *only*-fillers.

[11] There were a few reasons for this move. The collective form with a plural subject, which we used in the previous experiment, should not contribute to a decrease in acceptability compared to other predicates with a singular subject, but its semantic effects are not fully understood. The construction 'John agrees with Mary that *p*' allows for a singular subject (and therefore simplifies issues of cumulative readings). It also has a better-understood presupposition (that Mary believes *p*), which allowed the use of a uniform underlying structure to construct the material (in which the first premise validates a presupposition of the second premise). Note that in a task testing the semantics of question-embedding *agree*, Chemla and George (2016) found no difference between these two constructions. They did not test embedded declaratives however.

[12] The reader may notice that our NM controls are actually not SNM (they were constructed with *only*, so they are SDE). However, since the 'fact' for these items did not validate the UE presupposition of *only*, the end result was an environment which made both directions of inference invalid (i.e., was non-monotonic). We preferred *only* over genuine SNM expressions such as 'exactly *n*', which were too hard to process to provide a proper baseline. The results confirmed our intuition that *only* is easily understood as non-monotonic.

Table 4 Structure for each item type appearing in Experiment 2

Predicate	Sentence	Schematic structure
Factive attitude *V*	Fact	The ADJ aliens $\text{PRED}_{\wedge}$.
	Strong/Weak	X *V* that the ADJ aliens $\text{PRED}_{+/-}$.
Agree	Fact	X_2 believes that the ADJ aliens $\text{PRED}_{\wedge}$.
	Strong/Weak	X_1 agrees with X_2 that the ADJ aliens $\text{PRED}_{+/-}$.
UE	Fact	There are ADJ_1 aliens and ADJ_2 aliens.
	Strong/Weak	The ADJ_1 aliens $\text{PRED}_{+/-}$.
DE	Fact	There are ADJ_1 aliens and ADJ_2 aliens.
	Strong/Weak	The ADJ_1 aliens didn't $\text{PRED}_{+/-}$.
NM	Fact	There are ADJ_1 aliens and ADJ_2 aliens.
	Strong/Weak	Only 12 ADJ_1 aliens $\text{PRED}_{+/-}$.
Only-fillers	Fact	There are ADJ_1 aliens and ADJ_2 aliens.
	Strong	Only 12 ADJ_1 aliens PRED_{+}.
	Weak	At least 5 ADJ_1 aliens PRED_{-}.

The 'fact' was always the first premise. The 'strong' and 'weak' sentences alternately played the roles of second premise and conclusion. All attitudes except *agree* were factive and shared the same structure, illustrated in the first line of the table. The *only*-fillers only appeared in a direct version (inference from the strong to the weak). See Appendix 2 for details about the lists from which proper names, ADJ, and the different types of PRED were drawn.

3.2.3 Participants

In all, 50 participants were recruited on Amazon's Mechanical Turk and were paid $2 for their participation (age range: 19–58, 29 males).

3.2.4 Statistical methods

For statistical analyses, responses were centered and normalized by participants. The responses to targets were further transformed, as discussed in Sect. 3.3.2 (we projected the two-dimensional response space <Direct, Indirect> onto two new dimensions determined by the responses to controls). As in the previous experiment, we used parsimonious mixed-effects models, following Bates et al. (2015a). We fitted random effects for PRED, which was the most crucial random variation between items (we did not fit random effects for ADJ since that would have increased the risk of models not converging and was unlikely to have an important effect to begin with).

3.3 Results

Figure 5 presents responses to Indirect items as a function of responses to Direct items, for each environment. In such a representation, SUE items should be attracted towards the lower-right corner and SDE items towards the opposite, upper-left corner. SNM items should fall away from this anti-diagonal and be attracted towards the lower-left corner.

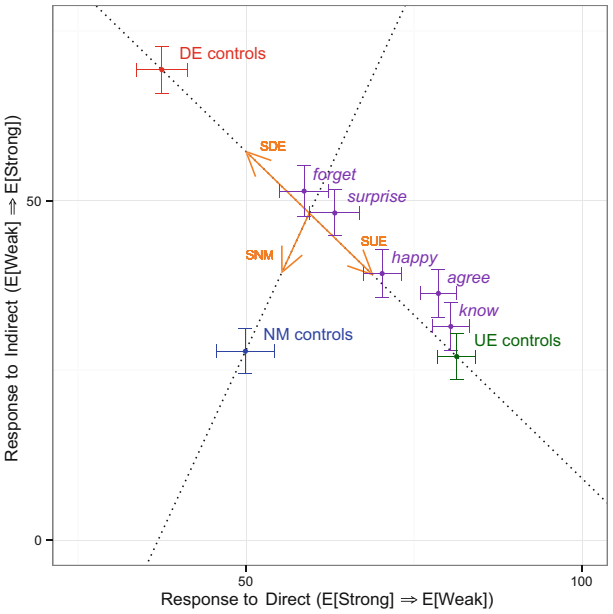

Fig. 5 Responses to Indirect as a function of responses to **Direct** items, for each environment. In such a representation, SUE predicates should fall to the bottom-right corner (making Direct inferences valid, but **Indirect** inferences invalid). SDE predicates are expected to reverse this pattern and appear at the top-left corner. SNM predicates should make any inference invalid, and therefore fall at the bottom-left corner. The two dashed lines represent the dimension which we projected on for the statistical analyses (although this projection was actually done on normalized responses rather than on the raw data). The orange arrows indicate the directions in which SUE, SDE, and SNM predicates should deviate from the central point.

3.3.1 Control items

We fitted a mixed-effects model for each inference type (Direct and Indirect) on the three control conditions with full random by-subject and by-item effects (Environment is both a within-subject and a within-item effect). The first one was fitted on responses to the Direct items; it showed that Direct inferences were judged more valid in UE controls than in NM controls ($t = 6.8$, $p < .001$), and more valid in NM controls than in DE controls ($t = 2.9$, $p = .004$). The Indirect inferences were judged equally low in UE and NM conditions ($t = -.2$, $p = .84$), but were significantly more valid in the DE condition ($t = 7.7$, $p < .001$).

3.3.2 Attitude predicates

Using the control items, we defined two new dimensions along which we analyzed the attitude predicates.

First dimension ("Deviation from monotonicity"): We determined the representative $\overline{UE}$, $\overline{DE}$, and $\overline{NM}$ positions in the bidimensional space of Direct and Indirect responses as the average of the responses to the UE, DE, and NM control items. We then computed the projection of each response on the axis starting at the midpoint of

the $[\overline{UE}, \overline{DE}]$ segment and going towards $\overline{NM}$. A non-null value on this axis represents a tendency towards non-monotonicity.[13] We then fitted a mixed-effects model on these dependent variables. We included Environment as a fixed effect (5 levels, corresponding to the 5 attitudes tested) and full random by-subject and by-item structure (including slopes for Environment, which as mentioned is both within-subject and within-item). None of the fixed effects was significant (all $|t| < .7$), and a model with only an intercept explained just as much variance ($\chi^2(4) = .6$, $p = .96$; note that the intercept itself was very close to zero $t = -.2$). This suggests that no predicate yields non-monotonicity.

Second dimension ("Upwardness"): We fitted a mixed-effects model on the projections on the $[\overline{UE}, \overline{DE}]$ line of responses for each predicate, to determine whether that verb was more SUE or SDE. This time, we found clear differences between verbs ($\chi^2(4) = 20$, $p < .001$). As a post-hoc analysis, we compared attitude predicates two-by-two, following their order on the $[\overline{UE}, \overline{DE}]$ line. Based on Bonferroni-corrected p-values for 4 comparisons, *know* was similar to *agree* ($\chi^2(1) = 3.4$, $p' = .26$), which was slightly more upward-monotonic than *be happy* ($\chi^2(1) = 5.6$, $p' = .073$). *Be happy* was clearly more upward monotonic than *surprise* ($\chi^2(1) = 11$, $p' = .003$), which did not differ significantly from *forget* ($\chi^2(1) = 1.3$, $p' = 1$). Even though *surprise* and *forget* were judged more downward entailing than other verbs, they were still very far from the DE controls.

3.4 Discussion

First, let us address a potential issue with complex presuppositions which we did not control for.[14] Some of the verbs we tested are not only factive but presuppose some form of knowledge: *forget* presupposes past knowledge, *surprise* presupposes that the agent came to know the complement, and *be happy* presupposes awareness. Provided basic assumptions about beliefs, these presuppositions are all upward-entailing. It seems natural to assume that participants accommodated these presuppositions in the premises, and therefore all presuppositions of the conclusion were satisfied in the direct targets (the x-axis in Fig. 5). However, this is not true of indirect targets. Indeed, even if participants accepted the premise "Benjamin forgot that the opaque aliens read books" and accommodated that Benjamin used to know that opaque aliens read books, this would not grant that the presupposition of "Benjamin forgot that the opaque aliens read sci-fi novels" is satisfied, and some participants could have rejected the indirect targets even though they have a SDE reading of *forget*, *be happy*, or *surprise*. In short, this issue could have led us to underestimate the values on the y-axis in Fig. 5. Therefore, the correct positions for *surprise*, *forget*, and *be happy* may be higher than our data suggest. But any point above the UE-DE line would represent a trivial environment (one that validates both UE and DE inferences) and these verbs are all exactly on the

[13] Negative values are not expected, since they would translate a tendency towards tautology (making too many inferences valid).

[14] Thanks to an anonymous reviewer for pointing this out.

line. We thus conclude that the upward-entailing presuppositions of the tested verbs did not affect the results.

Most verbs ended up being perceived as upward entailing. We may imagine that this is an artifact of the task, which could induce a general "upwardness" bias. This may be due to low-level strategies adopted by some of the participants. A few of them reported basing their answers on the subcategory/super-category relations between complements of the embedded clauses. These kinds of strategies would ultimately lead a participant to ignore the embedding environment and respond based on the complement clauses alone (creating a de facto UE environment). Nevertheless, participants as a group made clear distinctions between the control items, in the expected directions: UE controls made the direct inference valid and the indirect inference invalid, DE controls reversed this pattern, and NM controls made all inferences less valid.

The results on attitude predicates, validated by the clear results on controls, thus suggest that we can distinguish two classes of predicates based on monotonicity: *know*, *agree*, and *be happy* on the one hand, which are clearly upward entailing, and *forget* and *surprise* on the other hand, which may be perceived as more downward entailing. We can draw two conclusions from this result. Crucially, the two emotive-factive predicates we tested were not perceived as less monotonic than other verbs. As a result, monotonicity appears to be orthogonal to the selectional properties of the verbs. Indeed, the groups that emerged here both contain verbs that embed *whether*-questions and verbs that do not.

No theory draws a direct connection between monotonicity properties and the acceptability of *whether*-questions. However, it has been suggested that (i) *whether*-questions are only acceptable under predicates for which a strongly exhaustive reading is available (Nicolae 2013, 2015; Guerzoni and Sharvit 2014), and (ii) that emotive-factive predicates do not allow strongly exhaustive readings because of their monotonicity properties (Uegaki 2015). If both (i) and (ii) were true, we would have observed a link between monotonicity properties and acceptability of *whether*-questions, but the results of Experiments 1 and 2 show a complete lack of correlation between these two properties. In the next experiment, we investigate the availability of strongly exhaustive readings under *know*, *forget*, and *surprise*, in order to understand which of (i) or (ii) is faulty (if not both).

4 Experiment 3: Strongly exhaustive readings

4.1 Motivations and additional background

The main goal of this experiment was to test the availability of strongly exhaustive (SE) readings for questions embedded under *know, forget*, and *surprise*. Berman (1991) was the first to argue that *surprise* only gives rise to weakly exhaustive (WE) readings, and most introspective judgments in the literature agree.[15]

[15] Among exceptions, some do not challenge Berman's judgment on the assertive component but argue for different presuppositions. In particular, Abels (2007) and George (2011) propose a mention-some reading

The (un)availability of SE readings with *surprise* has been linked to the monotonicity properties of this verb, as well as, independently, to its selectional properties. Importantly, however, *forget* shares the monotonicity properties of *surprise* (Experiment 2), but patterns with *know* when it comes to selectional properties (Experiment 1), so if there is a link between these two properties, it has to be mediated by yet another factor. By testing the availability of SE readings with these three verbs, we should in principle be able to disentangle all possible links between monotonicity, exhaustivity, and selectional properties.

Before presenting our experiment, let us first review basic assumptions about *know*, *forget*, and *surprise*. Our goal is to list conceivable readings for sentences where they embed questions. We will then be in a position to evaluate the availability of each of these readings. More specifically, we will be able to assess whether *surprise* gives rise to an SE reading, factoring out the influence of other possible readings.

4.1.1 Know

Denotation: We will assume a very simple denotation for *know*: 'John knows p' presupposes that p is true and asserts that John believes p. There are constraints on what types of beliefs count as (justified) knowledge, but we designed our experiment such that the beliefs attributed to agents clearly satisfy these constraints.

Readings: Know has received more attention than any other question-embedding predicate, and this is reflected in the wide variety of readings proposed in the literature. We will provide a list of these readings in (19).

Different *exhaustive* readings have been proposed, which have in common that for John to know who was at the party, they at least require that for each person who actually was at the party, John know that they were. The WE reading does not require anything else (Karttunen 1977), whereas the Intermediate Exhaustive (IE) reading further requires that John would not falsely believe that anyone else was at the party (Groenendijk and Stokhof 1982; Berman 1991; Preuss 2001; Spector 2005), and the most stringent Strong Exhaustive (SE) reading requires John to know that no one else was at the party (Groenendijk and Stokhof 1982).

It has been argued that questions embedded under *know* can sometimes give rise to weaker, non-exhaustive readings, which do not require the agent to know a complete answer to the question. As an example, (18) may be true as soon as Rupert knows of one place where Italian newspapers can be purchased, without knowing the exhaustive list of all such places. This reading is often called the *mention-some* reading.

(18) Rupert knows where he can buy an Italian newspaper.

Footnote 15 continued
instead of the WE reading (hence a weaker presupposition), while Spector and Egré (2015) propose a WE reading for the assertion, but an SE presupposition. Although everyone acknowledges that the WE reading is available, Klinedinst and Rothschild (2011, fn. 18) argue that an SE reading is possible in some circumstances, and Theiler (2014) elaborates on this idea by suggesting that emotive-factives are ambiguous between a *literal* reading which only gives rise to WE readings and a *deductive* reading which leads to SE readings.

As for exhaustive readings, different mention-some readings have been proposed in the literature. In particular, George (2011, 2013) argues for a stronger notion of mention-some reading, which requires Rupert to know one true answer to the question, but also not to have any false beliefs regarding the selling of Italian newspaper at other places. We will call this the 'false-belief-sensitive mention-some' reading, which we will abbreviate as FBS-MS. By contrast, we will call 'weak mention-some' (WMS) the mention-some reading that requires knowledge of at least one true answer without any further constraint.

Cremers and Chemla (2016) and Xiang (2016, Chap. 4) provide experimental evidence for the availability of all exhaustive readings (WE, IE, and SE); Phillips and George (2016) and Xiang (Xiang (2016), Chap. 4) do so for the mention-some readings (WMS and FBS-MS). This leads us to the list of readings in (19).

(19) Mary knows which of her cards are spades.

WE:	For each actual spade, Mary knows that it is a spade.
IE:	For each actual spade, Mary knows that it is a spade, and she does not falsely believe that any other card is a spade.
SE:	For each actual spade, Mary knows that it is a spade, and she knows that no other card is a spade.
WMS:	There is an actual spade that Mary knows to be a spade.
FBS-MS:	There is an actual spade that Mary knows to be a spade, and she has no false beliefs regarding other cards.

Let us make a note about presuppositions. Although *know* normally triggers a factive presupposition, no such presupposition arises with embedded questions. Instead, the different readings all relate agents to propositions which are actually true (true answers), and a factive presupposition does not show up beyond this (because, in short, it would simply say that the true answers are true).

4.1.2 Forget

Denotation: Little has been said about *forget*, but the semi-formal denotation in (20) should not be controversial. According to (20), *forget* presupposes that there used to be a correct belief (which is different from a factive presupposition, because it is in the past), and asserts that there is no such belief anymore.[16]

(20) $[\![\text{forget}]\!] = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believed}]\!](p)(x)};\ \neg[\![\text{believe}]\!](p)(x)\right]$

Readings: Fewer readings have been proposed for *forget*. Most theories predict that it only receives a WE reading. Heim (1994) proposed a systematic way to strengthen WE into SE readings, which applies here to complete the list in (21).[17]

[16] We simplify details about time, and simply assume that the presupposition is about a time strictly anterior to the assertion.

[17] There is no systematic definition of an IE reading which would apply blindly to the embedding predicate. Instead, there could be several ways to generalize the IE reading to verbs other than *know* (e.g., adding absence of false belief to the WE reading, or adding negation of a presupposition-less version of the verb),

(21) Mary forgot which of her cards are spades.

WE: Mary does not remember that all actual spades are spades.
= For at least one actual spade, she forgot that it is a spade.
SE: Mary does not remember that all actual spades are spades and all other cards are not spades.
= For at least one card, she forgot whether it is a spade.

Two remarks are in order. First, *forget* presupposes past knowledge, and as a result the WE reading is predicted to presuppose that Mary used to know which of her cards are spades in the WE sense, and the SE reading that she used to know it in the SE sense. In our experiment, we ensured that there was always support for the strongest possible presupposition, so that this should not play any role. Specifically, the design ensured that the agent had known at some point the suit of each card, a fact which is stronger than any possible presupposition for the relevant sentence. Second, we leave aside issues regarding homogeneity effects among the entities being quantified over in the embedded questions (Cremers, in prep; Križ 2015), which may give rise to stronger readings (e.g., 'Mary forgot of *all* actual spades that they are spades'). To anticipate on the results, it turns out such effects did not play any role.

4.1.3 Surprise

Denotation: Several denotations have been proposed for *surprise*. Most prominently, Guerzoni and Sharvit (2007) proposed the Strawson-downward-entailing entry in (22) (from which we dropped the presupposition of speaker-factivity, which is not relevant for our purposes). According to this denotation, 'x was surprised that p' presupposes that p is true and that x knows it, and it asserts that x expected $\neg p$ to be true. We may also consider the weaker denotation in (23), with which the assertion would merely be that x did not expect p to be true. We refer to this denotation as the 'NE' denotation, for 'Not Expected'. We note that the wide scope of the negation here is closer to what we see in the denotation we considered for *forget* in (20).

(22) $[\![\text{surprise}]\!]_{\text{G\&S}} = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believe}]\!](p)(x)};\ [\![\text{expected}]\!](\neg p)(x)\right]$

(23) $[\![\text{surprise}]\!]_{\text{NE}} = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believe}]\!](p)(x)};\ \neg[\![\text{expected}]\!](p)(x)\right]$

Uegaki (2015) proposed yet another entry, given in (24), which is non-monotonic. Given the results from Experiment 2, which showed no sign of non-monotonicity for *surprise*, we disregard this denotation.

(24) $[\![\text{surprise}]\!]_{\text{U}} = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believe}]\!](p)(x)};\ \forall w' \in DOX_x^w,\ Sim_{w'}(\neg p) <^{\text{exp}}_{x,w} Sim_{w'}(p)\right]$
$\approx$ 'x expected $\neg p$ more than she expected p'

Footnote 17 continued
but in practice, for *forget* and downward-entailing predicates in general, all theories block what could correspond to such an IE reading.

Readings: The WE and SE readings corresponding to each of the denotations in (22) and (23) are given in (25). There is a debate, however, as to how the presupposition of *surprise* surfaces when it embeds questions (Abels 2007; George 2011; Spector and Egré 2015). In the experiment described below, we abstracted away from these difficulties by making sure that the strongest possible presupposition was supported by the evaluation context, as we did for *forget*. Specifically, the setup ensured that the agent eventually knows the suit of each card.

(25) Jacob was surprised by which of his cards are clubs.

WE:	Jacob expected some actual clubs not to be clubs.
SE:	Jacob expected some actual clubs not be clubs, or some other cards to be clubs.
NE-WE:	It is not the case that Jacob expected all actual clubs to be clubs.
NE-SE:	It is not the case that Jacob expected all actual clubs to be clubs and all other cards not to be clubs.

Having surveyed the possible readings for sentences with embedded questions under *know* (19), *forget* (21), and *surprise* (25), we will now present our experiment on the availability of these readings, including the controversial SE reading for *surprise*.

4.2 Methods

4.2.1 Task

We adapted the truth-value judgment task that Cremers and Chemla (2016) used to detect exhaustive readings for questions embedded under *know*. The task requires participants to evaluate the truth of sentences with embedded questions against a general background (given in the instructions) and a picture representing both a situation in the actual world and the mental state of the subject of the sentence in this situation. The relevant mental states were beliefs for *know* and *forget* and expectations for *surprise*. Different situations then introduced various discrepancies between the agent's mental state and the presented reality, which made different readings of the relevant sentences true and false. This allowed us to run a reading detection analysis, as explained below.

4.2.2 Instructions and training phase

At the beginning of the experiment, participants received the following instructions:

(26) "A group of friends is playing a kind of poker game in which each player gets dealt a hand of 7 cards. When they receive their cards, they only have a quick look at them and try not to show any emotion, while still remembering the cards.

You will see the hands that one of them got in each round, and either which suits they expected or what they remember about the cards. The players often mistake the suits (for instance they remember clubs instead of spades), and sometimes they completely forget what some of their cards were.

You will have to judge whether sentences about their memories or expectations are true or false."

As we mentioned in the Introduction, all tested verbs have a factive presupposition which is trivially satisfied when they embed questions, but *forget* and *surprise* have additional presuppositions: *forget* presupposes that the agent used to know the answer to the question, while *surprise* presupposes that she now knows it. While there may be some debates regarding how much the agent had to know/must have come to know, the story in the instructions ensured that the agent saw all of his/her cards at some point, and therefore ensured that the context would support even the strongest presupposition.

The instructions also included 4 example items with explanations about why the correct answer was True or False in each case. After the instructions, participants were presented with 5 training items (in random order) to help them familiarize with the task. After answering a training item, participants received feedback which consisted in a green "Correct!" message for 350ms if the answer was correct, and a red message for 8s containing explanations on why their answer was incorrect otherwise. The disparity in feedback durations created an incentive to be accurate and allowed participants to look back at the picture to understand their mistake when they gave a wrong answer.

4.2.3 Design and stimuli

Each item consisted of two rows of seven playing cards and a sentence that participants were asked to judge true or false, as can be seen in the examples presented in Fig. 6. Stimuli were based on two crossed factors: Predicate (3 levels), which was pertinent (mostly) for the sentence, and Situation type (5 levels), which was pertinent for the picture.

Sentences: The sentences were all of the form "NAME VERB which of [his/her] cards are SUIT", where NAME was drawn randomly from a list of 20 male and 20 female names, VERB was one of 'knows', 'forgot', or 'was surprised by', the possessive pronoun agreed with the name, and SUIT was one of 'hearts', 'diamonds', 'clubs', and 'spades'. For instance, a sentence for *forget* could be "Mary forgot which of her cards are clubs". Our goal was to explore the truth-conditions of the sentences by displaying them in various conditions, introduced through pictures.

Pictures: Pictures displayed two rows of playing cards. The first row was meant to represent the 'actual' situation. The second row represented the relevant mental state of an agent about each of the cards in the first row. For *know* and *forget*, the relevant type of mental state is belief, and each card in the second row could thus be a match with the first row (correct belief), a mismatch (false belief), or a special type of card representing absence of belief by displaying two shapes simultaneously. For *surprise*, the relevant type of mental state concerns expectations, and the card could thus be a match (satisfied expectation), a mismatch (failed expectation), or a special card representing absence of firm expectation by representing two shapes simultaneously.

This card system thus allows one to represent different types of relations between mental states of an agent and the actual world. In particular, one can introduce two types of discrepancies: I(gnorance) or C(onflict) about a particular card. And the discrepancy could concern either a card which is actually of the suit mentioned in the sentence

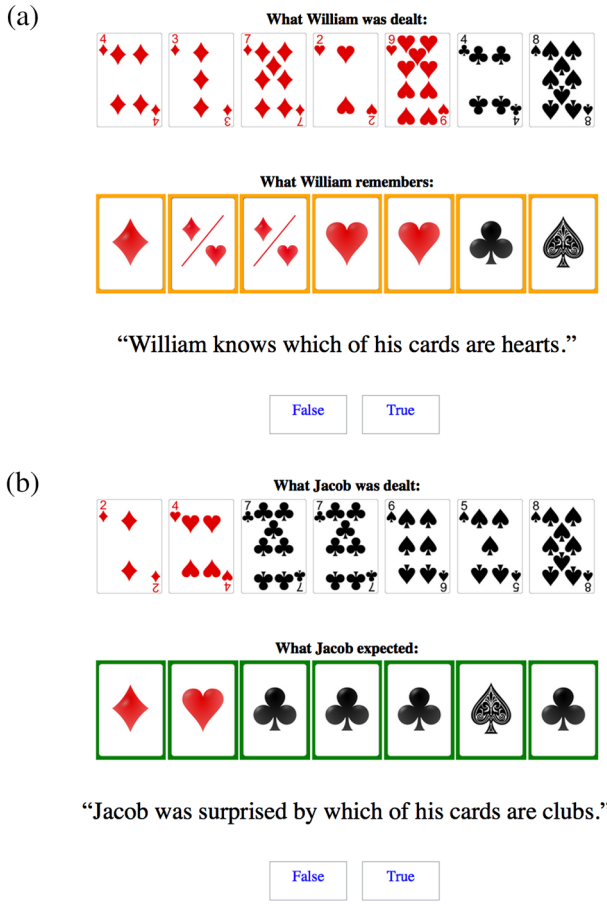

Fig. 6 Two items from Experiment 3. (a) An $I^{\ominus}$ target for *know*. Under an SE reading, we expect a 'False' answer. Under a WE or an IE reading, we expect a 'True' answer. (b) A $C^{\ominus}$ target for *surprise*. Under an SE reading, we expect a 'True' answer. Under a WE reading, we expect a 'False' answer.

($\oplus$), or a card which is actually part of a suit different than the one mentioned in the sentence ($\ominus$). Very much in the spirit of Xiang (2016, Chap. 4) then, we constructed five conditions: one with no discrepancy at all (N), and four with discrepancies of each of the different 2×2 possible types, $I^{\oplus}$, $I^{\ominus}$, $C^{\oplus}$, $C^{\ominus}$, which—for *know* and *forget*—correspond to the under-affirming, under-denying, over-denying, and over-affirming conditions respectively. For *surprise*, $I^{\oplus}$ amounts to the expectation that at most the actual target cards, and possibly less, be of the correct suit; $I^{\oplus}$ amounts to the expectation that at least the actual cards, and possibly more, be of the target suit; and $C^{\oplus}$ and $C^{\ominus}$ amount to the expectation of strictly less and strictly more cards of the target suit, respectively. When the condition involved one of these discrepancies, that discrepancy was always instantiated by two cards.[18]

[18] More situations could in principle be constructed and tested: some which would mix different types of discrepancies (e.g., an actual false belief on a positive card and ignorance on a negative card, a $C^{\oplus}$+$I^{\ominus}$

All cards were taken randomly from the range of ace to 10 (thus excluding the king, queen, and jack, which might have been harder to categorize). The actual cards were always sorted in the following order: diamonds, hearts, clubs, spades. Depending on whether the sentence was about a red suit (heart/diamond) or a black suit (club/spade), there were 5 red cards and 2 black cards, or the opposite. The expectations/memories described by the test sentences were only about the suits, not about the pips (numbers); therefore cards in the second row did not have a pip but only a suit symbol (this also discouraged purely visual matching of the top and bottom cards). Whenever there was a discrepancy between the two rows, it was always within color (i.e., a player could mistake a heart for a diamond, but not for a spade).[19]

Interaction between sentences and pictures: Table 5 provides the predicted truth value of each reading of each sentence in each condition. The table makes visible a few facts we discussed earlier. First, the WE, IE, and SE readings only diverge in the "negative" situations $I^{\ominus}$ and $C^{\ominus}$, i.e. cases where the discrepancies between a player's mental state and the actual world concern cards which are not of the target suit (the "negative" part of the predicate). Second, we can see what responses are predicted. If we assume the 'not-expected' denotation for *surprise* in (23), we obtain the exact same predictions as for *forget*. If we assume the more standard denotation for *surprise* in (22), we then obtain different predictions in the Ignorance conditions ($\ominus$ or $\oplus$): If Chris has no specific expectation regarding which suit a card will be, we predict that he cannot be surprised when it turns out to be a diamond (with *forget*, if Chris has no idea whether his third card is a heart or a diamond, we *can* say he forgot the suit of this card).

4.2.4 Participants

In all, 47 participants were recruited on Amazon's Mechanical Turk and were paid $1.80 for their participation (age range: 19–65). Two participants were excluded from the analysis because they did not report English as their native language. Three more participants were excluded because their error rate exceeded the average by more than one standard deviation (threshold: 34%; error rates were calculated on uncontroversial items for which all readings lead to the same truth value in Table 5).

Footnote 18 continued
situation, say) or some which would cover the cards differently (e.g., contrasting violated expectations on *some* positive cards vs. violated expectations on *all* positive cards). These possibilities were left aside for the time being. The 'simple' situations were sufficient to test all the readings listed in (19), (21), and (25).

[19] This, and the fact that Ignorance was represented as a hesitation between two suits of the same color, helped control for possible non-monotonic results of *surprise*, corresponding to the denotation proposed by Uegaki (2015). Indeed, he suggests that $[\![\text{surprise}]\!](p)(x) = 1$ if and only iff x expected $\neg p$ more than p. If we were to represent Ignorance with a question mark as in Cremers and Chemla (2016), there would be four alternatives (corresponding to the four suits). Assuming they are all equally likely, this would yield a probability of $3/4$ for $\neg p$ ('the card is not a club') and only $1/4$ for p ('the card is a club'), thus making Ignorance practically equivalent to Conflict. Here we made sure there were only two alternatives, ensuring equal probability to p and $\neg p$.

Table 5 Readings tested for *know*, *forget*, and *surprise*, the expected responses for a participant who would access each of these readings, and the estimated proportions of responses in our data that can be attributed to each reading

(a)

Reading	N	$I^{\ominus}$	$C^{\ominus}$	$I^{\oplus}$	$C^{\oplus}$	Confidence interval	Statistics
SE	1	0	0	0	0	38% – 88%	$\chi^2(1) = 37, \quad p < .001$
IE	1	1	0	0	0	<10%	$\chi^2(1) = .1, \quad p = .721$
WE	1	1	1	0	0	8% – 20%	$\chi^2(1) = 29, \quad p < .001$
FBS-MS	1	1	0	1	0	10% – 24%	$\chi^2(1) = 43, \quad p < .001$
WMS	1	1	1	1	1	2% – 6%	Could not be tested

(b)

Reading	N	$I^{\ominus}$	$C^{\ominus}$	$I^{\oplus}$	$C^{\oplus}$	Confidence interval	Statistics
SE	0	1	1	1	1	56% – 85%	$\chi^2(1) = 115, \quad p < .001$
WE	0	0	0	1	1	9% – 29%	$\chi^2(1) = 24, \quad p < .001$

(c)

Reading	N	$I^{\ominus}$	$C^{\ominus}$	$I^{\oplus}$	$C^{\oplus}$	Confidence interval	Statistics
SE	0	0	1	0	1	8% – 55%	$\chi^2(1) = 17, \quad p < .001$
NE-SE	0	1	1	1	1	4% – 48%	$\chi^2(1) = 7.6, \quad p = .006$
WE	0	0	0	0	1	5% – 29%	$\chi^2(1) = 15, \quad p < .001$
NE-WE	0	0	0	1	1	?	Could not be tested

(a) Readings, predicted responses, and estimated share of participants' responses for each reading of "X knows which of his/her cards were clubs". The proportion of WMS readings could not be statistically tested because we lacked a false baseline to compare the $C^{\oplus}$ condition to.
(b) Readings, predicted responses, and estimated share of participants' responses for each reading of "X forgot which of his/her cards were clubs". Due to the limited number of readings proposed for *forget*, this model could not explain any difference between the $I^{\ominus}$ and $C^{\ominus}$ conditions, nor between the $I^{\oplus}$ and $C^{\oplus}$ conditions.
(c) Readings, predicted responses, and estimated share of participants' responses for each reading of "X was surprised by which of his/her cards were clubs".

4.2.5 Analytical and statistical methods

The results were submitted to a reading detection analysis (which has a precursor in Chemla and Spector 2014). We optimized models of the response data, using as predictors the possible readings (as listed in (19) and (25)). Hence, a positive coefficient in such a model is evidence for the availability of a particular reading (when other plausible readings are taken into account). For each parameter (i.e., each reading), we give the statistics obtained from model comparison as well as a confidence interval in

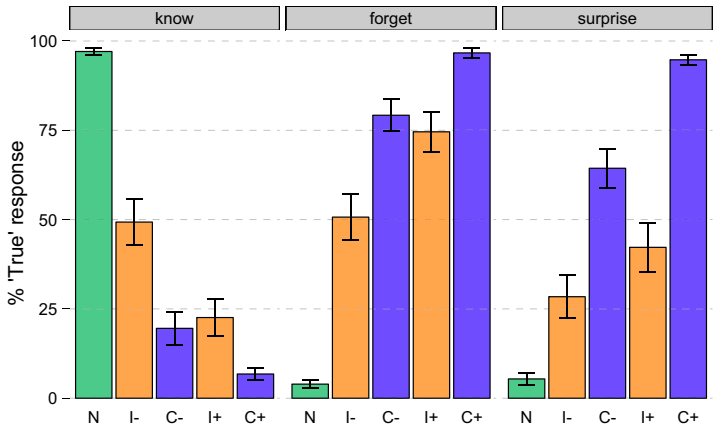

Fig. 7 Percentage of 'True' responses for each verb and condition in Experiment 3. In a 2-by-2 comparison, the only differences which turned out non-significant were between the $C^{\ominus}$ and $I^{\oplus}$ conditions under *know* and *forget*, and between the $I^{\ominus}$ and $I^{\oplus}$ conditions under *surprise*

percent,[20] which we interpret as an approximation of the proportion of responses that can be attributed to the corresponding reading.

The fastest and slowest 2% of responses were removed. Remaining responses were analyzed with mixed-effects logit models with random effects for participants and items (encoded by suits). Because logit mixed models are computationally more demanding than the linear mixed models we used in previous experiments, we often had to drop several random effects before obtaining models such that at least the parsimonious models derived from them would converge (Bates et al. 2015a). We tried to fit maximal random-effects structures for participants, but only random intercepts for items (the effects were usually very small for items).

4.3 Results

Figure 7 presents the raw responses for each condition and for each verb.[21] Table 5 presents the results of the statistical analyses using readings as predictors as described above, and the corresponding estimation of the role of each reading in terms of explained proportion of responses for *know* and *surprise*.

We provide more detailed analyses for each of the verbs below, but we can immediately summarize the results. For *know*, we obtain clear evidence for SE readings, WE readings, and FBS-MS readings. The results for *forget* are best understood as the mirror image of the results for *know*, as if the readings of *forget* were the exact negations of the readings of *know* (more on this below). For *surprise*, there was clear

[20] These confidence intervals were obtained by estimating confidence intervals on the logit scale, transposing them by the estimated intercept, and converting back to the 0–100% scale.

[21] As the figure suggests, we note that all simple two-by-two comparisons between conditions within each verb turned out significant ($\chi^2(1) > 5$), except for the $C^{\ominus}$ and $I^{\oplus}$ conditions on *know* and *forget* items and the $I^{\oplus}$ and $I^{\ominus}$ conditions on *surprise* items (all three $\chi^2(1) < 1.3$).

evidence for the WE reading, as expected, as well as for the more controversial SE and NE-SE readings.

4.3.1 Know

We fitted a logit mixed model on responses to *know* items with predictors corresponding to the five readings in (19). Note that the WMS reading is true in all five conditions; it therefore corresponded to the intercept of the model, i.e., the rate of True answers when all predictors/other readings are false, which also happens to correspond to the $C^{\oplus}$ condition. In light of this, and in the absence of a baseline for errors, we could compute a confidence interval but no informative p-value.

The results, reported in Table 5a, indicate that all readings except the IE reading were significantly detected. As suggested by the low rate of True responses in the $C^{\oplus}$ condition, the WMS reading was estimated below 6% and may still be conflated with simple errors.

4.3.2 Forget

We ran a similar model for *forget*. As indicated in Table 5b, both potential readings (WE and SE) played a significant role. Yet, looking at Fig. 7, it is clear that this model does not provide a satisfying explanation of the results on *forget* items. Due to the very limited number of readings proposed for questions embedded under *forget*, it cannot explain all the differences we observed, e.g., between the $I^{\ominus}$ and $C^{\ominus}$ conditions, or between the $I^{\oplus}$ and $C^{\oplus}$ conditions.

A simple look at the raw data in Fig. 7 shows that *know* and *forget* behaved like a True/False mirror image of each other: responses to *forget* were strongly anti-correlated to responses to *know*, across conditions and participants (Pearson's product-moment correlation: $\rho = -.94$, $t(208) = -39$, $p < .001$).

We ran a logit mixed model on responses to *know* and *forget*, after flipping all responses to *forget* items (encoding True responses as False and *vice versa*, in order to directly compare the rates of True responses to *know* with the rates of False responses to *forget*). The model included Condition (5 levels), Verb (2 levels) and their interactions as fixed effects, and random slopes for Verb per participant, but not for Condition (it would not converge otherwise), and random intercepts for Items (suits). Overall, the factor Verb explained very little variance ($\chi^2(5) = 7.2$, $p = .20$); judging from the estimated z-values, only one interaction corresponding to the $C^{\oplus}$ condition approached significance, and only did so before correction for multiple comparison ($z = 2.2$, $p = .03$, all other z's $< .5$).

4.3.3 Surprise

On responses to *surprise* items, we fitted a model with the 4 readings in (25) as predictors, as well as an intercept. Given that none of the tested readings was true in the P condition, the intercept roughly corresponded to the P condition and could be interpreted as the baseline rate of True responses due to errors. Dropping either the WE or the NE-WE reading had little effect on the model's fit (both $\chi^2(1) < 1.6$, $p > .2$),

yet dropping both had a very significant effect ($\chi^2(2) = 19, p < .001$). It turned out that the two predictors were strongly correlated (estimated correlation of the fixed effects: $-.85$).

Going back to the raw data, as previously mentioned, there was no significant difference between the $I^{\oplus}$ and $I^{\ominus}$ conditions, which only differ on the NE-WE reading ($\chi^2(1) = .2, p = .6$), but there was a major difference between the $C^{\oplus}$ and $C^{\ominus}$ conditions, which differ on the WE and NE-WE readings ($\chi^2(1) = 9, p = .003$). We therefore decided to drop the NE-WE reading from the model. In the second model, reported in Table 5c, all three remaining readings were significantly detected (all $\chi^2(1) > 7, p < .01$).

4.4 Discussion

4.4.1 SE readings for all predicates

To sum up our results: First, we partially replicated Cremers and Chemla's (2016) results about *know*, in that we observed WE and SE readings. We also found a significant proportion of FBS-MS readings (which have the no-false-belief constraint of the IE reading without the exhaustivity). However, we did not obtain evidence of the IE reading, which was the predominant reading found in Cremers and Chemla (2016).[22]

The results on *forget* are somehow puzzling. One might think that, provided the presuppositions of both verbs are satisfied, the proposition-embedding *forget* would be equivalent to the negation of *know*. If we assume the question-embedding entries to be reducible to the proposition-embedding entries, we would predict the WE readings of *know* and *forget* to be negations of each other (and possibly so for their SE readings). However, this would *not* explain why we observe a reading of *forget* which seems to be the negation of the FBS-MS reading of *know* (because this reading is crucially a case of non-reducibility; see George 2011, 2013).[23] The most plausible interpretation of the result for *forget*, then, is a task-specific strategy, according to which participants deal with the *forget* sentences by first evaluating corresponding, simpler sentences with *know* and then flip their answers. Such an effect may for instance be reduced in a between-subject design.

For *surprise*, we detected WE and SE readings. We also detected the NE-SE reading, which suggests that some participants felt that an agent could be said to be surprised even when she did not exclude that the actual outcome was possible. Since this NE-SE reading does not seem to correspond to the intuitive meaning of *surprise*, it is

[22] The current task elicits more SE readings for *know* than Cremers and Chemla's (2016) task. Note that if this was due to a bias towards False answers, the chances to detect SE readings with *surprise* would be reduced.

[23] The only existing theories we are aware of which derive the FBS-MS reading are Theiler et al. (2016) and Xiang (2015). However, there is a general agreement that its derivation should parallel that of the IE reading on the exhaustive side. Yet, no theory predicts IE readings for *forget*. For the theories in the exhaustification tradition (Klinedinst and Rothschild 2011; Uegaki 2015), the IE reading is the result of a pragmatic strengthening of the WE reading and this strengthening is vacuous when applied to *forget*. For alternative theories (Spector and Egré 2015; Theiler et al. 2016), the IE reading is obtained through a different composition rule or operator, which is blocked for downward-entailing predicates.

tempting to attribute this result to a low-level strategy again. Some possibilities come to mind. For instance, one could try to defend the idea that participants come to accept sentences with *surprise* in all conditions but N, or to treat *surprise* as a negation of *know* with a strongly exhaustive reading. Yet another, more sophisticated possibility is that (i) participants relied on the non-monotonic reading in (24), and (ii) in a case of absence of expectations, they assumed that the agent had stronger expectations for the wrong suit. This second assumption basically reduces the NE-SE to a genuine SE reading with Uegaki's denotation (24). Thus we might hypothesize that we found a clear SE reading (with the standard denotation) and then maybe, on top of this, other SE readings with a 'not-expected' denotation or with Uegaki's denotation (plus auxiliary, task-related assumptions). In the next section, we will focus on the SE reading associated with the most standard denotation, as evidenced by the very high proportion of True responses in the $C^{\ominus}$ condition, and discuss in more detail possible alternative explanations of this important fact.

4.4.2 SE readings with surprise and alternative interpretations

We interpret the high rate of True responses in the $C^{\ominus}$ condition as evidence that *surprise* can give rise to SE readings. This conclusion goes against a long history of introspective judgments. An alternative view on this result has been suggested to us.[24] Imagine that participants interpreted or read "Jacob was surprised by which of his cards are clubs" as (27). It seems intuitive that the latter is true in the $C^{\ominus}$ condition, which would thus explain the high rate of True responses in that condition. But if this is correct, it would still be evidence in favor of an SE reading here, according to Beck and Rullmann (1999), who predict the readings in (28).

(27) Jacob was surprised by how many of his cards are clubs.

(28) a. WE: Jacob expected a lower number of clubs among his cards.
b. SE: Jacob expected a different number of clubs among his cards.

Hence, even under this reinterpretation view, we obtain evidence for an SE reading of questions embedded under *surprise*, albeit maybe for *how-many* rather than *which* questions.[25] In principle, we could have avoided this issue entirely by replacing cards

[24] Thanks to the audience at SIASSI Berlin 2015, and in particular Angelika Kratzer, for this suggestion. We would like to point out that the objection would apply equally to our experimental results and to the example presented by Klinedinst and Rothschild (2011, fn. 18).

[25] One may wonder whether the reading described as an SE reading in (28b) can be obtained differently, and in particular by directly assuming an exact reading for numerals (as opposed to an at-least reading, as in Beck and Rullmann 1999). This is in fact possible, but then we would not be able to obtain the other reading, (28a). A strong argument for a theory that can derive the (28a) reading of *how-many* questions comes from the asymmetry between (i) and (ii), already noted by Beck and Rullmann (1999). With exact readings only, this asymmetry cannot be captured.

(i) It surprised Mary how many guests showed up.
Interpretation: It surprised her that *so many* guests showed up.

(ii) It surprised Mary how many eggs are sufficient to bake this cake.
Interpretation: It surprised her that *so few* eggs are needed.

with entities which are well individuated. For instance, when talking about guests sitting at a table, it may be easier to distinguish expectations regarding each individual guest from a more general expectation regarding the total number of guests who showed up.[26]

Finally, one could imagine other reinterpretations of the embedded question. The logic behind the *how-many* reinterpretation was that in the context of a card game it may not matter which cards are clubs, but simply how many of each suit you got. We showed that even under this reinterpretation, the True responses were underlyingly SE readings, but this may not be the case in general. We will not discuss this broader objection because it is too vague to be rejected. Indeed, with no limit on which reinterpretations must be considered, one can always build a modified sentence the WE reading of which corresponds to the SE reading of the original sentence (for instance, "It surprised Mary which cards are clubs [and which are not]").

4.5 Summary for Experiment 3

In Experiment 3, we found that all verbs we tested gave rise to SE readings. While the interpretation of the results for *forget* is obscured by an apparent low-level strategy of the participants, we are confident that the results with *surprise* do indicate genuine SE readings. The availability of SE readings is a challenge for many theories which are either designed to derive the unavailability of this reading (Guerzoni and Sharvit 2007; Romero in press) or take it as the cause for the unacceptability of *whether*-questions (Nicolae 2013, 2015; Guerzoni and Sharvit 2014).

5 Conclusion

The acceptability judgments collected in Experiment 1 confirmed the soundness of the typology proposed in the literature (in line with previous studies on the convergence of quantitative methods and introspective judgements; Sprouse and Almeida 2013 and Sprouse et al. 2013). Each class of verbs did indeed allow embedding of the type of complements they were expected to allow, and there was little variation within each class. In particular, we confirmed the widespread introspective judgment regarding the unacceptability of *whether*-questions under emotive-factive predicates. However, this unacceptability seems to be 'soft', in that *whether*-questions are not as unacceptable under emotive-factives as they are under anti-rogative verbs (in line with Sæbø 2007). We also found that the acceptability of an adverb of quantity follows Beck and Sharvit's (2002) judgments on the availability of QVE, in particular with rogative predicates, and discovered an interesting contrast between polar and alternative questions which seems to reflect dialectal variation within American English and may support Guerzoni and Sharvit's (2014) derivation of polar questions as alternative questions.

Footnote 25 continued
Summing up, there is strong evidence for at-least readings of *how-many* questions. The most straightforward explanation for the possibility of the reading in (28b) then is that it is an SE reading (and not an independently existing exact reading for *how-many* questions).

[26] Thanks to Andreea Nicolae for this suggestion.

Experiment 2 collected inferential judgments between sentences with various responsive predicates embedding declarative complements. In this experiment, *forget* and *surprise* patterned together, and were judged to be more downward-entailing than *know*, *agree*, and *be happy*. These results suggest that monotonicity is independent from selectional properties, since the distinction between upward and downward monotonicity is orthogonal to the typology confirmed in Experiment 1. All responsive predicates were judged monotonic.

Experiment 3 was a truth-value judgment task. A reading detection analysis revealed that both *know* and *surprise* give rise to SE readings when embedding questions. The fact that *surprise* does give rise to SE readings goes against most introspective judgments in the literature. It also suggests that the unacceptability of *whether*-questions under emotive-factives cannot be explained by the unavailability of an SE reading.

Although no theory has directly linked selectional properties of emotive-factive predicates with their monotonicity, it has been argued that their selectional properties may be explained by the unavailability of SE readings, which in turn may be explained by monotonicity properties. The results of Experiments 1 and 2 showed that monotonicity and selectional properties vary independently, so the two connections cannot both hold. The results of Experiments 1 and 3 suggest that the broken link is (at least) the connection between strong exhaustivity and the acceptability of *whether*-questions (contra Nicolae 2013, 2015; Guerzoni and Sharvit 2014). Given the evidence, strong exhaustivity may still be linked to monotonicity (or perceived monotonicity; Chemla et al. 2011), however. Indeed, *surprise* was not judged to be as downward-entailing as the DE controls in Experiment 2, and it may be that the participants who had an SE reading for *surprise* in Experiment 3 did not access a clearly Strawson-downward-entailing meaning for this verb.

In sum, the present studies provide a range of empirical facts about emotive-factives and other question-embedding predicates, describing acceptable constructions, monotonicity properties, and potential readings. These studies immediately contribute to some empirical debates in the literature, in particular by providing the strongest evidence to date in favor of the existence of SE readings for questions embedded under *surprise*. Two broad options are available for future research on this topic; both of these were suggested to us independently by Andreea Nicolae and Yael Sharvit. First, following Chemla et al. (2011), it would be interesting to run all the experiments we presented with a single set of participants, in order to test for direct correlations between one's acceptance of *whether*-questions, perceived Strawson-monotonicity, and choice of an exhaustive reading for embedded questions. Second, it would be interesting to investigate the distribution of NPIs under emotive-factive predicates and see if it correlates with any of these three factors.

Appendix 1: Lists for Experiment 1

Our set of proper names consisted of 20 of the most frequent male and female names in the U.S. We selected 28 adjectives which described appearance or physical properties for the aliens, to avoid interaction with the activity predicates as much as possible.

Names: *Alex, Anthony, Ashley, Ava, Benjamin, Chloe, Chris, Elizabeth, Emily, Emma, Ethan, Grace, Jacob, James, John, Madison, Michael, Olivia, Sarah, William*

ADJ: *blue, cubical, flexible, fluffy, fuzzy, green, hollow, large, opaque, orange, pink, purple, red, round, slender, slim, small, smooth, speckled, spherical, spiky, spiny, spotty, stringy, striped, transparent, watery, yellow*

For activity predicates, we created the 20 predicates in the PRED column of Table 6. We also used versions with a disjunction to generate alternative questions, in the $\text{PRED}_{\vee}$ column.

Table 6 List of PRED used in Experiment 1

PRED	$\text{PRED}_{\vee}$
Play the piano with their wings	Play the piano with their wings or with their feet
Can eat 5 pounds of licorice	Can eat 5 pounds or 10 pounds of licorice
Drink soda with a spoon	Drink soda with a straw or with a spoon
Write poems about the moon	Write poems about the moon or about the sun
Ride tall purple horses	Ride purple horses or blue horses
Read 18th century books	Read 18th century or 19th century books
Visit archeology museums	Visit archeology museums or geology museums
Sleep with their head down	Sleep with their head up or with their head down
Lay translucent eggs	Lay translucent eggs or opaque eggs
Speak several African languages	Speak Asian languages or African languages
Can count to two million	Can count to two million or to three million
Freeze certain tropical plants	Freeze or fry certain tropical plants
Talk with their noses	Talk with their noses or with their ears
Hibernate every three years	Hibernate every three years or twice a year
Believe in green unicorns	Believe in green unicorns or blue dragons
Listen to classical music	Listen to classical music or to techno music
Watch movies from the 30's	Watch movies from the 30's or from the 40's
Use smartphones to cook pasta	Use smartphones to cook pasta or to serve wine
Drive old cars	Drive old cars or rusty motorbikes
Use the Korean alphabet	Use the Korean alphabet or the Greek alphabet

Appendix 2: Lists for Experiment 2

We used the same proper names as in Experiment 1. We used the same 28 adjectives, but each was paired with another related adjective in case it would appear in one of the control conditions (which required two different adjectives). The pairing is presented in Table 7. The predicates used in this experiment are presented in Table 8, in their weak (PRED_{-}), strong (PRED_{+}), and conjunctive ($\text{PRED}_{\wedge}$) versions.

Table 7 List of ADJ used in Experiment 2

ADJ_1	ADJ_2	(continued)	
Blue	Red	Slender	Slim
Cubical	Spherical	Slim	Stringy
Flexible	Rigid	Small	Large
Fluffy	Spiky	Smooth	Fuzzy
Fuzzy	Fluffy	Speckled	Slender
Green	Purple	Spherical	Cubical
Hollow	Watery	Spiky	Spiny
Large	Small	Spiny	Speckled
Opaque	Transparent	Spotty	Striped
Orange	Blue	Stringy	Round
Pink	Orange	Striped	Spotty
Purple	Yellow	Transparent	Opaque
Red	Green	Watery	Hollow
Round	Flexible	Yellow	Pink

Table 8 List of weak, strong, and conjunctive versions of PRED used in Experiment 2

$PRED_-$	$PRED_+$	$PRED_\wedge$
Burn flowers	Burn roses	Burned roses and tulips
Buy clothes	Buy shirts	Bought shirts and trousers
Color trees	Color pines	Colored pines and oaks
Compliment humans	Compliment children	Complimented children and teenagers
Destroy musical instruments	Destroy violins	Destroyed violins and guitars
Drink sodas	Drink coke	Drank coke and lemonade
Drive cars	Drive Toyotas	Drove Toyotas and Fords
Eat at restaurants	Eat at Mexican restaurants	Ate at Mexican and Chinese restaurants
Eat meat	Eat pork	Ate pork and beef
Kiss animals	Kiss dogs	Kissed dogs and cats
Play with toys	Play with toy cars	Played with toy cars and toy soldiers
Read books	Read sci-fi novels	Read sci-fi novels and love novels
Read magazines	Read news magazines	Read news magazines and sports magazines
See birds	See doves	Saw doves and crows
Taste cookies	Taste chocolate cookies	Tasted chocolate cookies and caramel cookies
Throw balls	Throw tennis balls	Threw tennis balls and soccer balls
Use coins	Use quarters	Used quarters and dimes
Use the internet	Send emails	Sent emails and visited websites
Visit museums	Visit French museums	Visited French museums and Italian museums
Watch sports matches	Watch baseball matches	Watched baseball matches and football matches

References

Abels, K. 2007. Deriving selectional properties of 'exclamative' predicates. In *Interface and interface conditions*, ed. A. Späth, 115–140. Berlin: De Gruyter.

Barr, D.J., R. Levy, C. Scheepers, and H.J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68 (3): 255–278.

Bates, D., R. Kliegl, S. Vasishth, and H. Baayen. 2015a. Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015b. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67 (1): 1–48.
Beck, S., and H. Rullmann. 1999. A flexible approach to exhaustivity in questions. *Natural Language Semantics* 7 (3): 249–298.
Beck, S., and Y. Sharvit. 2002. Pluralities of questions. *Journal of Semantics* 19 (2): 105–157.
Berman, S. R. 1991. On the semantics and logical form of wh-clauses. PhD thesis, University of Massachusetts, Amherst.
Chemla, E. 2009. Presuppositions of quantified sentences: Experimental data. *Natural Language Semantics* 17 (4): 299–340.
Chemla, E., and B.R. George. 2016. Can we agree about agree? *Review of Philosophy and Psychology* 7 (1): 243–264.
Chemla, E., and B. Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28: 359–400.
Chemla, E., and B. Spector. 2014. Distinguishing typicality and ambiguities: the case of local scalar implicatures. Manuscript, LSCP and IJN.
Chemla, E., V. Homer, and D. Rothschild. 2011. Modularity and intuitions in formal semantics: The case of polarity items. *Linguistics and Philosophy* 34 (6): 537–570.
Core Team, R. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
Cremers, A. (in prep). Plurality effects and exhaustivity with embedded questions. Manuscript, Universiteit van Amsterdam.
Cremers, A., and E. Chemla. 2016. A psycholinguistic study of the exhaustive readings of embedded questions. *Journal of Semantics* 33 (1): 49–85.
Egré, P. 2008. Question-embedding and factivity. *Grazer Philosophische Studien* 77 (1): 85–125.
George, B. R. 2011. Question embedding and the semantics of answers. PhD thesis, University of California Los Angeles.
George, B.R. 2013. Knowing-'wh', mention-some readings, and non-reducibility. *Thought: A Journal of Philosophy* 2 (2): 166–177.
Geurts, B., and F. van der Slik. 2005. Monotonicity and processing load. *Journal of Semantics* 22 (1): 97–117.
Grice, P. 1975. Logic and conversation. In *The Logic of Grammar*, ed. D. Davidson and G.H. Harman, 64–75. Encino, CA: Dickenson Publishing Company.
Grimshaw, J. 1979. Complement selection and the lexicon. *Linguistic Inquiry* 10 (2): 279–326.
Groenendijk, J., and M. Stokhof. 1982. Semantic analysis of wh-complements. *Linguistics and Philosophy* 5 (2): 175–233.
Groenendijk, J., and M. Stokhof. 1984. Studies on the semantics of questions and the pragmatics of answers. PhD thesis, University of Amsterdam.
Guerzoni, E. 2003. Why even ask? On the pragmatics of questions and the semantics of answers. PhD thesis, MIT.
Guerzoni, E. 2007. Weak exhaustivity and 'whether': A pragmatic approach. In *Proceedings of SALT 17*, 112–119. Washington, DC: LSA.
Guerzoni, E., and Y. Sharvit. 2007. A question of strength: on NPIs in interrogative clauses. *Linguistics and Philosophy* 30 (3): 361–391.
Guerzoni, E., and Y. Sharvit. 2014. *Whether or not anything but not whether anything or not*. In *The Art and Craft of Semantics: A Festschrift for Irene Heim*, vol. 1, ed. L. Crnic and U. Sauerland, 199–224. Cambridge, MA: MIT Working Papers in Linguistics.
Heim, I. 1994. Interrogative semantics and Karttunen's semantics for "know". In *IATL 1*, vol. 1, ed. R. Buchalla and A. Mittwoch, 128–144, Jerusalem: Hebrew University.
Herbstritt, M. 2014. Why can't we be surprised whether it rains in Amsterdam? A semantics for factive verbs and embedded questions. Master's thesis, Universiteit van Amsterdam.
Karttunen, L. 1977. Syntax and semantics of questions. *Linguistics and Philosophy* 1 (1): 3–44.
Klinedinst, N., and D. Rothschild. 2011. Exhaustivity in questions with non-factives. *Semantics and Pragmatics* 4 (2): 1–23.
Križ, M. 2015. Homogeneity, trivalence and embedded questions. In *Proceedings of the 20th Amsterdam Colloquium*, ed. T. Brochhagen, F. Roelofsen, and N. Theiler, 207–216. University of Amsterdam.
Lahiri, U. (2002). *Questions and answers in embedded contexts* (Oxford studies in theoretical linguistics 2). New York: Oxford University Press.

Nicolae, A. C. 2013. Any questions? Polarity as a window into the structure of questions. PhD thesis, Harvard.
Nicolae, A.C. 2015. Questions with NPIs. *Natural Language Semantics* 23 (1): 21–76.
Phillips, J., and B.R. George. 2016. Non-reducibility with knowledge wh: Experimental investigations. Manuscript, Harvard and Carnegie Mellon University.
Preuss, S.M.-L. 2001. Issues in the semantics of questions with quantifiers. PhD thesis, Rutgers.
Roelofsen, F., M. Herbstritt, and M. Aloni. (To appear). The **whether* puzzle. In *Questions in discourse*, ed. K. von Heusinger, E. Onea, and M. Zimmermann. Leiden: Brill.
Romero, M. 2015. Surprise-predicates, strong exhaustivity and alternative questions. In *Proceedings of SALT 25*, ed. S. D'Antonio, M. Moroney, and C.R. Little, 225–245. Washington, DC: LSA.
Sæbø, K.J. 2007. A whether forecast. In *Logic, Language and Computation*, ed. B. ten Cate and H. Zeevat, 189–199. Berlin: Springer.
Spector, B. 2005. Exhaustive interpretations: What to say and what not to say. Presentation at LSA Workshop 'Context and Content', July 15, 2005.
Spector, B., and P. Egré. 2015. Embedded questions revisited: An answer, not necessarily the answer. *Synthese* 192 (6): 1729–1784.
Sprouse, J., and D. Almeida. 2013. The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes* 28 (3): 222–228.
Sprouse, J., C.T. Schütze, and D. Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134: 219–248.
Theiler, N. 2014. A multitude of answers: Embedded questions in typed inquisitive semantics. Master's thesis, ILLC, University of Amsterdam.
Theiler, N., F. Roelofsen, and M. Aloni. 2016. A truthful resolution semantics for declarative and interrogative complements. Manuscript, ILLC, University of Amsterdam.
Uegaki, W. 2015. Interpreting questions under attitudes. PhD thesis, MIT.
Wickham, H. 2009. *ggplot2: Elegant graphics for data analysis*. New York: Springer.
Wilkinson, K. 1996. The scope of even. *Natural Language Semantics* 4 (3): 193–215.
Xiang, Y. 2015. Complete and true: A uniform analysis for mention-some and mention-all questions. In *Proceedings of Sinn und Bedeutung 20*, ed. N. Bade, P. Berezovskaya, and A. Schöller, 815–832. Tübingen University.
Xiang, Y. 2016. Interpreting questions with non-exhaustive answers. PhD thesis, Harvard.